Research Scholar Details	
Name Number	Mr. Gaurav Omprakash Gajbhiye
Mobile	+91-8149946140
Email id	gauravogajbhiye@gmail.com, gajbhiyegaurav@sggs.ac.in
Category	SC
Subject	Electronics and Telecommunication Engineering
Faculty	Science and Technology
Title/Research topic	Understanding and Generating Natural Language Description for Digital Images
Date of RAC	14/09/17
Date of RRC	28/03/18
Institute Registration Number	2017PEC604
Research Center	Shri Guru Gobind Singhji Institute of Engineering and Technology, Vishnupuri, Nanded
Permanent address	Yashoda Nagar No.1, Sawalkar Market, Behind Adaneshwar Temple, Amravati, Maharashtra. 44606
Residential address	C/o Mr. Kokare, Amrut Residency Flat no. 307, Third Floor, Near SGGSIE&T, Vishnupuri, Nanded.
Research Guide/Supervisor Details	
Name	Dr. Abhijeet V. Nandedkar
Mobile Number	+91-9970933062
Email id	avnandedkar@gmail.com, avnandedkar@sggs.ac.in
Designation	Associate Prof., Dept. of Electronics and Telecommunication Engg.
Institute	Shri Guru Gobind Singhji Institute of Engineering and Technology, Vishnupuri, Nanded

Synopsis for Registration as Research Scholar

for

Doctoral Degree

on

Understanding and Generating Natural Language Description for Digital Images

By

Gajbhiye Gaurav Omprakash

Research Scholar

Supervisor

Dr. Abhijeet V. Nandedkar

Department of Electronics and Telecommunication Engineering

Shri Guru Gobind Singhji Institute of Engineering and Technology,

Nanded-431606 (M.S.) India.

Understanding and Generating Natural Language Description for Digital Images

Abstract:

The rapid increase in the exchange of digital visual data with its description in text on the internet requires a strong relationship between the two modalities for conveying the proper information. Exact picture information need to be extracted for a given sentences and a proper sentence need to be generated for given pictorial data. This vigorous problem can be solved by combining two stages of image understanding and natural language generation. An intelligent model proposed which simultaneously learns to create the feature vector for image and for its corresponding description. A model is smart enough to learn and map both image and its description. Image feature computed by fully connected layer of Convolutional Neural Network (CNN) and associated sentences are converted to feature vector by Long Short Time Memory (LSTM). Recurrent Neural Network (RNN) with LSTM build a superior model to learn and train large number of images with its sentences. For any test image, according to image features given by CNN, LSTM use previous knowledge along with the probability of next word and decide to generate best possible word. Model recurrently work in same fashion to provide appropriate sentence for a given image.

1. Introduction:

The development of the Artificial Intelligence (AI) demands recognition of objects in visual imagery and is providing a superb results. But description of the content in image is still remains a challenging task. Several Machine Learning (ML) techniques viz. object detection, classification and recognition gives specific information about the content in the image, but it will not provide the semantic knowledge of imagery. Great deal of work has concentrated on expanding new techniques that automatically generates the sentence based on picture information [1]. Machines can learn to understand the image with its description and hence able to generate new description for relevant images [2]. Development of fast computing ability of machine and invention of new deep Neural Network algorithm helps AI researcher to figure out the solution for difficult AI task such as image description [3, 4, 5]. Integrating visual content of imagery data with the language learning algorithm for description has been considered as a new thrust of research for solving different types of problems. Automatic translation of the imagery information into the linguistic sentences with natural language processing is the new AI task termed as Image Captioning [7, 8]. The standard approach for image captioning are image understanding through Computer Vision (CV) techniques and sentence generation through Natural Language Processing (NLP) techniques. Image captioning is fundamental problem in AI and becomes immerging trend for various application as Image/Video retrieval through linguistic description, Image/Video Understanding, Blind Navigation and automatic image description [10].

Many pioneering approaches has been proposed for solving the AI task of Image Captioning. There are three main approaches used by various researches to find the solution of sentence generation

for given image as template based method, semantic transfer based method, and neural network based methods [7].

In template based method [9], initially objects, attributes and relationship among these objects in image detected. Corresponding words related to detected objects, attributes and relationship filled into rigid sentence template to generate the complete sentence. This method requires tremendous data to train the model but not having enough flexibility for correct sentence generation.

In transfer based method [8], According to query image, relevant image is retrieved from the annotated image data set. The description of retrieved image is transferred to query image. As compared to template based method, transfer based method generates human like linguistic sentence but it require huge amount of annotated data set to get better performance.

In neural network based method [4, 5, 6] deep neural network algorithm with encoder-decoder pipeline have been used for image understanding and sentence generation. Recently, this kind of method achieved great success in image\video description. This method has more flexibility and can generate close human like linguistic sentence. Most of the researcher still working for development of algorithm to succeed in image to sentence translation. This synopsis also concentrate on the various methodology in deep neural network to translate image into its corresponding linguistic sentence.

The principle behind the image captioning are based on supervised learning of images with its annotation. Data base for image captioning requires the plenty of images with its linguistic description. Most of the researcher validate the proposed approach on standard datasets viz. Flickr8k, Flickr30k, MSCOCO. These data base contains large amount of image and images were annotated by language experts [4, 5, 6, 7]. The standard approach to image description endorse a two-way framework consisting of content selection and surface realization. First stage investigate different objects in the image and determines “what to say” (i.e. which event or objects are worth talking about), whereas second stage builds a language model which describes “how to say” (i.e. how to render the selected content into natural language text) [1]. Generally in image captioning, Deep Convolutional Neural Network (CNN) employed for content selection and Recurrent Neural Network (RNN) with Long Short Term Memory (LSTM) employed for generation of linguistic sentences [7, 9, 10]. By combining the modified version of CNN, RNN and LSTM, Automatic caption generation achieves highest degree of result [4, 5, 6, 7, 9].

Supervised convolutional neural network (CNN) model works smoothly for image recognition, object classification [4]. CNN can produce a rich representation of the input images by embedding it to a fixed-length vector [6]. Multiple layers viz. convolutional layer, activation layer, max-pooling layer included in NN to build the CNN which learn high level image feature with better discrimination and robustness. Convolutional layer in CNN extracts local image pattern and invariant to translation. An activation layer designed to add non-linearity to network and to enhance its representation ability. Popular activation function employed in the activation layer are sigmoid, tanh and ReLU (Rectified Linear Unit). “max-pooling” operation carried out after

activation layer to determine corresponding maximum value of local area of input image. Finally output of multiple layers of max-pooling concatenate to get Fully Connected Layer or fixed length vector which classifies object in image [9]. Recurrent neural network (RNN) basically use to learn sequential data pattern. An input sequence feed to "encoder" RNN which translates into fixed-length vector and this vector employed as hidden layer for "decoder" RNN that generates target [6]. In RNN, "Vanishing Gradient" problem occurs due to back propagation of error through long term dynamics. Long Short Term Memory (LSTM) proposed to overcome vanishing gradient and provide the exceptional facility to network. Network able to learn when to escape previous hidden state and when to update previous hidden state in memory unit [4, 10]. LSTM have ability to large scale of data patterns [10].

In general, image captioning can be achieved by first finding the various objects and attributes in image. Then neural network model is learned to map the objects regions and attributes with corresponding words in a sentences. Natural language generator designed to infer the words for objects and attributes and then arrange it in a proper sequence to produce appropriate natural language sentence.

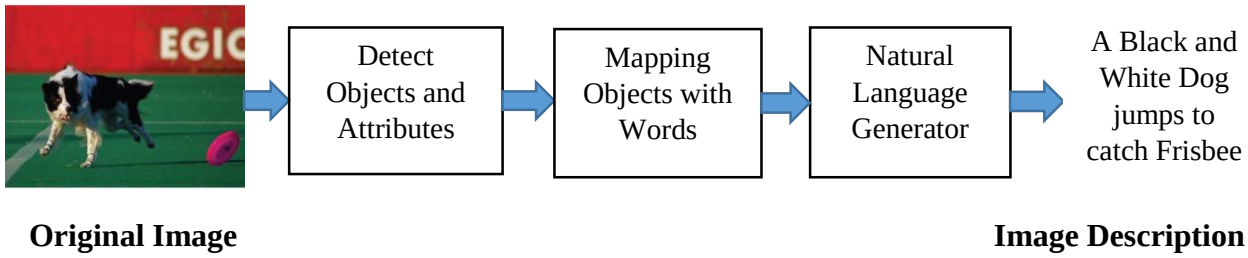
Detect Objects and Attributes
Mapping Objects with Words
Natural Language Generator
A Black and White Dog jumps to catch Frisbee
Original Image
Image Description

Fig 1. General image captioning methodology

2. Literature Review:

From last few years, there is extensive growth of digital information on the internet. Due to vast availability and accessibility of machine, peoples upload and download the huge amount of imaginary data with its descriptions through internet. Most of work succeeded for image understanding using various Computer Vision approaches and recently, more focused carrying on automatic sentence generation for unknown image. Most of the researchers have developed new AI techniques for automatic caption generation.

Initially, news images with its article is utilized for caption generation. An intelligent model consisting of content selection and surface realization created to learn captions from the database containing news article with picture embedding in it [1]. Latent Dirichlet Allocation (LDA) train on multimodal document containing image-caption-document tuple which represents bag of words. LDA learn multimodal topic representation and use inference to obtain topic distributions of unseen image document pairs [1]. It becomes difficult for machine to learn from captioning as many intermediate texts cannot be defined by objects. This can be overcome by determining the

objects in sub regions of images by rich convolutional neural network (CNN) features. Features of sub region are mapped with the words and train this map by applying multiple instance learning (MIL) which learns different image for each text word. Global deep multimodal similarity model introduced to re-rank the captions [2]. New framework designed for annotation prediction and caption extraction by incorporating the contextual information of image which connects the semantic gap between image and words [3]. Framework accept the news as input consisting of image, auxiliary information source, article, news category label and keywords to produce sentence like caption for image. Framework compute the probability of all four attribute for test image and predict the best word annotation by combining the context probability of all four sources [3].

Kulkarni G. et *al.* [4] has given introductory model for understanding and generating simple image description. Author proposed system in which detector are used to detect the objects and stuff. Then candidate region processed by set of attribute classifiers. Prepositional relationship function is employed for selection of proper preposition. Conditional Random field (CRF) is employed to anticipate best label for input image. Meaningful sentence is generated from labeling.

Activity Recognition, Image Captioning and Video Description highly achieved with great efficiency by Long Term Recurrent Convolutional Network with deep LSTM. Donahue J. et *al.* [5] proposed LSTM approach in RNN followed by CNN with input sequence of image and its description. A Model proposed by A. Karpathy et *al.* [6] based on the novel approach of combination of convolutional neural network (CNN) over image regions, bidirectional recurrent neural network (BRNN) over sentences and a structured objective that align these two modalities. Multimodal embedding structure links image and sentence features to generate natural language description. Similar type of work proposed by O. Vinyals et *al*. [7] in which CNN is used as image "encoder" for producing fixed length vector and RNN is used with LSTM as "decoder" for producing natural language sentence corresponding to image. Keeping the same algorithm for "encoder" and making certain changes in "decoder", Kaisheng Xu et *al*. [8] proposed a deep residual LSTM (ResLSTM) language model for image captioning. Temporal dropout technique utilized in proposed framework to increase robustness of network.

One of the pioneering research in transfer based methodology for data driven image captioning is given by Kilickaya et *al.* [9]. Combination of deep visual feature and object based representation employed to retrieve relevant images from massive set of images. Both scene and object information is use to transfer the caption for query image. A novel phrase selection approach utilized by [9] in which several captions collected from retrieved image and appropriate caption transfer for sentence generation. Shi et *al.* [10] proposed a new template based framework for generation of linguistic description for remote sensing image. Fully Convolutional Layer (FCN) used for image understanding and corresponding sentence generated by using template based methodology. Different sentence template designed by arranging the triplets of ELM (refers to ground elements), ATR (refers to attribute such as quantity, size and location) and RLT (refers to relationship of different elements).

Some more advancement has done by Gao et *al.* [11] for video captioning by utilizing attention based LSTM. Attention mechanism adopts dynamic weighted sum of local two-dimensional CNN representation. LSTM decoder integrates image feature at time 't' and word embedding feature at time 't-1' to provoke intelligent wording. Finally multimodal embedding employed to align visual and wording features into joint space to ensure the semantic consistence of sentence description and the video visual content. A complex methodology for image captioning promoted by Fu K. et *al.* [12]. Author has extracted visual feature by dividing the whole image into multiple visual regions. Visual feature provided to LSTM to learn the sequence of text data with visual region. A global visual context called scene vector also extracted from the whole image. Attention based LSTM decoder is used to give fine description of image.

3. Objective of Research and Proposed Hypothesis:

Image Captioning is fundamental task in Artificial Intelligence which describes objects, scene and relationship of these objects in image with the human-like sentences. It has gain instant interest in the field of research and currently large amount of work is carrying throughout the world. However, Image Captioning is quite difficult task but improvement in GPU (Graphics Processing Unit) and Advance algorithms of Neural Network helps to achieve high degree of efficiency and to design various framework for image captioning. The major objectives of image captioning can be classified in three main category as:

1. Describe the visual information of image in linguistic mode.
2. Establish the relationship between scenery knowledge in image and natural language sentence.
3. Express the semantic knowledge of image in natural language like English.

To achieve above objectives, this work is planned as follows:

1. Detection of shape, size and color of objects in image.
2. Detection of attributes and scenery (action) knowledge in image.
3. Identifying the relation between objects in image.
4. Build a model to map the objects, attributes and their relation with corresponding text words.
5. Introduce the natural language algorithm to learn the sentence for images.
6. Generate the human like sentence for test image to describe content and scenery knowledge of image.

4. Methodology to be adopted:

Natural image captioning methodology basically divided into two stages: 1) Image Understanding and 2) Natural Language Generation. This work will focus on generalized methodology for image

understanding and natural language generation. This work will solve the problem of Image Captioning by Supervised learning algorithm, in which model firstly train with images and its annotation. Corresponding caption is generated by maximum probability distribution for test image. Following methodology will be adopted to achieve the objectives.

a) Data Base:

- We are going to collect a database containing huge amount of images with its annotation from various resources to build a model.

b) Training Phase:

- For training, images with its descriptions is given to multimodal space model.
- One of the strongest class of deep neural network, Convolutional Neural Network (CNN) employed on image for classification of different objects.
- A fixed length feature vector is generated from output of CNN for given input image.
- At the same time, Recurrent Neural Network (RNN) will transform the words in sentences to the vector of same length as in CNN.
- RNN is not sufficient to learn the large sequence of words, new model Long Short Term Memory (LSTM) is embed with RNN to learn and generate new sentence.
- Set of vector generated from CNN and RNN mapped in common multimodal space and find high matching score for individual word to sub-region of image.
- Large amount of images with its descriptions embed into multimodal space model to train and map the each word with its corresponding regions in image.

c) Testing Phase:

- To predict the caption for given test image, initially compute the image representation using CNN and set first token of word as "START".
- According to the objects & attributes in images, RNN will decide the next word by considering previous hidden state value and maximum probability distribution of words in a stack.
- LSTM will help for generating next word by deciding which previous word need to be forget and which word need to be updated in a cell state.
- Procedure repeats until last token of word obtain as "END".
- A sensible sentence generated by repeating these steps.

5. Importance of study/ Society application:

- **International Status:**

Automatic caption generation for images is currently hot issue and excellent work carried out to solve this complex AI task throughout the world. Many pioneering research

conducted to promote new methodology for image to sentence translation. Some of the latest researches enlisted here with.

Tran K. et *al.* [13] proposed a three models viz. deep vision model for detection of visual concepts, recognition model to identify celebrities and landmarks, and confidence model for caption generation. Deep residual convolutional NN used in vision and recognition model. Maximum entropy language model (MELM) with deep multimodal similarity model (DMSM) used for sentence generation.

Image captioning and visual question answering (VQA) achieved by Wu Qi. et *al.* [14] by providing attributes and external knowledge about the image to the proposed model. A supervised CNN model is used to detect the set of attributes in an image and corresponding words found in caption to train deep CNN model by minimizing element wise logistic loss. A fixed length vector having size of attributes is created for each image. Each dimension of vector provides the probability of particular attribute. LSTM employed with vector for generation of caption. Due external knowledge of attributes, Model provide more fined caption for image and also can be used for Question and Answering.

You Q. et *al.* [15] proposes a semantic attention model which is a combination of Top-to-bottom (image to word conversion) and Bottom-to-top (words describing image attributes) approach for image captioning.

A crucial breakthrough given by Shin Hoo-chang et *al.* [20] in the field of automatic medical image annotation. A model is proposed for learning to read chest X-ray images and gives annotation for given X-ray images. CNN algorithm applied for image feature calculation and RNN with combination of LSTM & GRU (Gated Recurrent Unit) applied to learn language feature. A combined Model of CNN-RNN (LSTM+GRU) provides a satisfied result of X-ray image annotation.

Globally various private organization such as Microsoft, Google, Facebook, Apple, Amazon, IBM etc. actively participating research in automatic image captioning. World's top university research group viz. Standford Vision Lab (Standford University), Barkeley Artificial Intelligence Research (University of California, Barkeley), Computer Science and Artificial Intelligence Laboratory (Massachusetts Institute of Technology) etc. engaged to provide a concrete solution for image caption for various application.

- **National Status**

  At national level, very few research papers are published on image captioning. Ghosh S. et *al.* [16] presented the model which will accept the image caption as input and retrieve the set of image depending on similarity of their caption with input test caption. The model basically deal with the natural language processing algorithm in which discourse information is extracted from input caption with the help of discourse representation structure. Directed attribute graph is used to convert the extracted information into data

structure which contain entity and relation for performing similarity analysis. Maximum similarity common subgraph is employed to calculate similarity score to return the set of images with similar kind of caption.

Stemming algorithm used by D. Ramya et *al.* [17] for generation of caption for news content images. Image and contents of news along with user keywords accepted by the model and stemming algorithm employed to reduce the inflected words from root words to give extractive content. Extractive contents compared with keywords and frequency ranking calculation is used to predict the caption for news image content.

Two stage framework consisting of content selection and surface realization proposed by Reddy A. S. et *al.* [18] for image captioning. Author basically used transfer based method in which, firstly visual features extracted from query image and annotated images of similar features return as output. Based on visual and semantic similarity, ranking algorithm employed to return exact caption for query image.

Fusion based recurrent multimodal learning proposed by Oruganti Ram M. et *al.* [19] for image description. Combination of three stages as input stage, language stage and fusion stage employed to design mechanism. During training, input stage extract the feature of image using fully connected layer of CNN and language stage consisting of LSTM learns the sentence structure, semantic and word occurrence probability from the ground truth image sentences. Image feature and language feature mapped into fusion stage to learn associated image features with language features. In testing phase, image feature extracted from input stage and corresponding sentence predicted by fusion stage by considering previous knowledge.

At national level, computer vision group of IIT Kanpur leads in research field image captioning and visual question answering. IIT Bombay has Center for Indian Language Technology which specially deals with natural language processing problems. Vision and Learning Research Group of Indian Statistical Institute has also gain research interest in the field of vision and language interpretation.

- **Significance of study:**

  In the era of Artificial Intelligence, Image captioning have vast practical usage which overcome plenty of down to earth problems related to Vision to Text translation, Image description, Image understanding, Retrieval of images using natural language sentences, etc. Some of the broad application of Image Captioning is given as below:

  1. **Captioning for News Images**: As news article consist of long explanation of news with its image. User's expectation to have a bold headline with image for better understanding of news.
  2. **Tag Based Image Retrieval:** Many social search engine works based on keywords for image retrieval. In such case it is difficult to access the images with exact group of

words. Image captioning help to serve this issue by retrieving the image with its annotation.

3. **Automatic Image Annotation:** Till now, people basically describe the image by their own hand. Image captioning becomes the pioneering technique to describe the image automatically.
4. **Medical Image Description:** Sometimes, it becomes difficult to diagnose a particular disease just by visually analyzing the medical image. A specialist may require all discernable information about the medical image in natural semantic wording. Image captioning helps to describe the medical image and specialist can easily diagnose the disease.
5. **Satellite Image Captioning:** It becomes necessary to detect and describe all the objects in satellite images to get knowledge of geographical aspects. Image captioning allows to give geographical knowledge in linguistic sentence for various satellite images.

6. Proposed Work Plan / Formulation and Structure of Study:

Year	Proposed Work Plan
2017-18	• Literature survey of image description and related work at international and national level • Study of machine learning, deep learning and AI related subject • Collection of Database • Learn python and other open source programming language • Implement base methods and verify the results • Overcome the lacuna in existing method by investigating new approach • Implement the new approach
2018-19	• Innovate the advancement in proposed approach to get better result. • Analyze and compare the proposed and existing outcomes. • Research Publication • Study the latest techniques to tackle the problem • Design a new approach with latest technique • Analyze and compare the result • Research Publication
2019-20	• Propose the advancement in architecture • Analyze and compare the result • Research Publication • Thesis Writing

References:

[1] Feng, Yansong, and Mirella Lapata. "Automatic caption generation for news images." *IEEE transactions on pattern analysis and machine intelligence* 35.4 (2013): 797-812.

[2] Fang, Hao, et *al.* "From captions to visual concepts and back." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

[3] Tariq, Amara, and Hassan Foroosh. "A context-driven extractive framework for generating realistic image descriptions." *IEEE Transactions on Image Processing* 26.2 (2017): 619-632.

[4] Kulkarni, Girish, et *al.* "Babytalk: Understanding and generating simple image descriptions." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.12 (2013): 2891-2903.

[5] Donahue, Jeffrey, et *al.* "Long-term recurrent convolutional networks for visual recognition and description." *IEEE transactions on pattern analysis and machine intelligence* 39.4 (2017): 677-691.

[6] Karpathy, Andrej, and Li Fei-Fei. "Deep visual-semantic alignments for generating image descriptions." *IEEE transactions on pattern analysis and machine intelligence* 39.4 (2017): 664-676.

[7] Vinyals, Oriol, et *al.* "Show and tell: Lessons learned from the 2015 mscoco image captioning challenge." *IEEE transactions on pattern analysis and machine intelligence* 39.4 (2017): 652-663.

[8] Xu, Kaisheng, Hanli Wang, and Pengjie Tang. "Image captioning with deep LSTM based on sequential residu*al.*" *Multimedia and Expo (ICME), 2017 IEEE International Conference on*. IEEE, 2017.

[9] Kilickaya, Mert, et *al.* "Data-driven image captioning via salient region discovery." *IET Computer Vision* (2017).

[10] Shi, Zhenwei, and Zhengxia Zou. "Can a Machine Generate Humanlike Language Descriptions for a Remote Sensing Image?" *IEEE Transactions on Geoscience and Remote Sensing*55.6 (2017): 3623-3634.

[11] Gao, Lianli, et *al.* "Video Captioning With Attention-Based LSTM and Semantic Consistency." *IEEE Transactions on Multimedia*19.9 (2017): 2045-2055.

[12] Fu, Kun, et *al.* "Aligning Where to See and What to Tell: Image Captioning with Region-based Attention and Scene-specific Contexts." *IEEE transactions on pattern analysis and machine intelligence* (2016).

[13] Tran, Kenneth, et *al.* "Rich image captioning in the wild." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2016.

[14] Wu, Qi, et *al.* "Image Captioning and Visual Question Answering Based on Attributes and External Knowledge." *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017).

[15] You, Quanzeng, et *al.* "Image captioning with semantic attention." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

[16] Ghosh, Swarnendu, et *al.* "Representing image captions as concept graphs using semantic information." *Advances in Computing, Communications and Informatics (ICACCI), 2016 International Conference on*. IEEE, 2016.

[17] Vijay, K., and D. Ramya. "Generation of caption selection for news images using stemming algorithm." *Computation of Power, Energy Information and Commuincation (ICCPEIC), 2015 International Conference on*. IEEE, 2015.

[18] Reddy, A. Sivakrishna, et *al.* "Automatic caption generation for annotated images by using clustering algorithm." *Innovations in Information, Embedded and Communication Systems (ICIIECS), 2015 International Conference on*. IEEE, 2015.

[19] Oruganti, Ram Manohar, et *al.* "Image description through fusion based recurrent multi-modal learning." *Image Processing (ICIP), 2016 IEEE International Conference on*. IEEE, 2016.

[20] Shin, Hoo-Chang, et *al.* "Learning to read chest X-rays: recurrent neural cascade model for automated image annotation." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

Gaurav O. Gajbhiye
Research Scholar

Dr. Abhijeet V. Nandedkar
Supervisor